Noise-Resilient Group Testing: Limitations and Constructions

Mahdi Cheraghchi*

EPFL, Switzerland

Abstract. We study combinatorial group testing schemes for learning d-sparse boolean vectors using highly unreliable disjunctive measurements. We consider an adversarial noise model that only limits the number of false observations, and show that any noise-resilient scheme in this model can only approximately reconstruct the sparse vector. On the positive side, we give a general framework for construction of highly noise-resilient group testing schemes using randomness condensers. Simple randomized instantiations of this construction give non-adaptive measurement schemes, with $m = O(d \log n)$ measurements, that allow efficient reconstruction of d-sparse vectors up to $O(d)$ false positives even in the presence of δm false positives and $\Omega(m/d)$ false negatives within the measurement outcomes, for *any* constant $\delta < 1$. None of these parameters can be substantially improved without dramatically affecting the others. Furthermore, we obtain several explicit (and incomparable) constructions, in particular one matching the randomized trade-off but using $m = O(d^{1+o(1)} \log n)$ measurements. We also obtain explicit constructions that allow fast reconstruction in time $\mathsf{poly}(m)$, which would be sublinear in n for sufficiently sparse vectors.

1 Introduction

Group testing is an area in applied combinatorics that deals with the following problem: Suppose that in a large population of individuals, it is suspected that a small number possess a condition or property that can only be certified by carrying out a particular test. Moreover suppose that a *pooling strategy* is permissible, namely, that it is possible to perform a test on a chosen group of individuals in parallel, in which case the outcome of the test would be positive if at least one of the individuals in the group possesses the condition. The trivial strategy would be to test each individual separately, which takes as many tests as the population size. The basic question in group testing is: how can we do better? This question is believed to be first posed by Dorfman [1] during the screening process of draftees in World War II. In this scenario, blood samples are drawn from a large number of people which are tested for a particular disease. If a number of samples are pooled in a group, on which the test

* Email: mahdi.cheraghchi@epfl.ch. Research supported by Swiss NSF grant 200020-115983/1.

is applied, the outcome would be positive if at least one of the samples in the group carries a particular antigen showing the disease. Since then, group testing has been applied for a wide range of purposes, from testing for defective items (e.g., defective light bulbs or resistors) as a part of industrial quality assurance [2] to DNA sequencing [3] and DNA library screening in molecular biology (see, e.g., [4,5,6,7,8] and the references therein), and less obvious applications such as multiaccess communication [9], data compression [10], pattern matching [11], streaming algorithms [12], software testing [13], and compressed sensing [14], to name a few. Moreover, over decades, a vast amount of tools and techniques has been developed for various settings of the problem that we cannot thoroughly survey here, due to space restrictions. Instead, we refer the reader to the books by Du and Hwang [15,16] for a detailed account of the major developments in this area.

More formally, the basic goal in group testing is to reconstruct a d-sparse[1] boolean vector[2] $x \in \mathbb{F}_2^n$, for a known integer parameter $d > 0$, from a set of observations. Each observation is the outcome of a measurement that outputs the bitwise OR of a prescribed subset of the coordinates in x. Hence, a measurement can be seen as a binary vector in $\mathbb{F}_2^n$ which is the characteristic vector of the subset of the coordinates being combined together. More generally, a set of m measurements can be seen as an $m \times n$ binary matrix (that we call the *measurement matrix*) whose rows define the individual measurements.

In this work we study group testing in presence of highly unreliable measurements that can produce false outcomes. We will mainly focus on situations where up to a constant fraction of the measurement outcomes can be incorrect. Moreover, we will mainly restrict our attention to *non-adaptive* measurements; the case in which the measurement matrix is fully determined before the observation outcomes are known. Nonadaptive measurements are particularly important for applications as they allow the tests to be performed independently and in parallel, which saves significant time and cost.

On the negative side, we show that when the measurements are allowed to be highly noisy, the original vector x cannot be uniquely reconstructed. Thus in this case it would be inevitable to resort to approximate reconstructions, i.e., producing a sparse vector $\hat{x}$ that is close to the original vector in Hamming distance. In particular, our result shows that if a constant fraction of the measurements can go wrong, the reconstruction might be different from the original vector in $\Omega(d)$ positions, irrespective of the number of measurements. For most applications this might be an unsatisfactory situation, as even a close estimate of the set of positives might not reveal whether any particular individual is defective or not, and in certain scenarios (such as an epidemic disease or industrial quality assurance) it is unacceptable to miss any affected individuals. This motivates us to focus on approximate reconstructions with *one-sided* error. Namely, we will

[1] We define a d-sparse vector as a vector whose number of nonzero coefficients is at most d.

[2] We use the notation $\mathbb{F}_q$ for a field of size q. Occasionally we adapt this notation to denote a set of size q as well even if we do not need the underlying field structure.

require that the support of $\hat{x}$ contains the support of x and be possibly larger by up to $O(d)$ positions. It can be argued that, for most applications, such a scheme is as good as exact reconstruction, as it allows one to significantly narrow-down the set of defectives to up to $O(d)$ *candidate positives*. In particular, as observed in [17], one can use a *second stage* if necessary and individually test the resulting set of candidates to identify the exact set of positives, hence resulting in a so-called *trivial two-stage* group testing algorithm. Next, we will show that in any scheme that produces no or little false negative in the reconstruction, only up to $O(1/d)$ fraction of false negatives (i.e., observation of a 0 instead of 1) in the measurements can be tolerated, while there is no such restriction on the amount of tolerable false positives. Thus, one-sided approximate reconstruction breaks down the symmetry between false positives and false negatives in our error model.

On the positive side, we give a general construction for noise-resilient measurement matrices that guarantees approximate reconstructions up to $O(d)$ false positives. Our main result is a general reduction from the noise-resilient group testing problem to construction of well-studied combinatorial objects known as *randomness condensers* that play an important role in theoretical computer science. Different qualities of the underlying condenser correspond to different qualities of the resulting group testing scheme, as we describe later. Using the state of the art in derandomization theory, we obtain different instantiations of our framework with incomparable properties summarized in Table 1. In particular, the resulting randomized constructions (obtained from optimal lossless condensers and extractors) can be set to tolerate (with overwhelming probability) *any* constant fraction (< 1) of false positives, an $\Omega(1/d)$ fraction of false negatives, and produce an accurate reconstruction up to $O(d)$ false positives (where the positive constant behind $O(\cdot)$ can be made arbitrarily small), which is the best trade-off one can hope for, all using only $O(d \log n)$ measurements. This almost matches the information-theoretic lower bound $\Omega(d \log(n/d))$ shown by simple counting. We will also show explicit (deterministic) constructions that can approach the optimal trade-off, and finally, those that are equipped with fully efficient reconstruction algorithms with running time polynomial in the number of measurements.

Related Work. There is a large body of work in the group testing literature that is related to the present work; in this short presentation, we are only able to discuss a few with the highest relevance. The exact group testing problem in the noiseless scenario is handled by what is known as *superimposed coding* (see [18,19]) or the closely related concepts of *cover-free families* or *disjunct matrices*[3]. It is known that, even for the noiseless case, exact reconstruction of d-sparse signals (when d is not too large) requires at least $\Omega(d^2 \log n / \log d)$

[3] A d-superimposed code is a collection of binary vectors with the property that from the bitwise OR of up to d words in the family one can uniquely identify the comprising vectors. A d-cover-free family is a collection of subsets of a universe, none of which is contained in any union of up to d of the other subsets. These notions are extended to the noisy setting, e.g., in [20].

measurements (several proofs of this fact are known, e.g., [21,22,23]). An important class of superimposed codes is constructed from combinatorial designs, among which we mention the construction based on MDS codes given by Kautz and Singleton [24], which, in the group testing notation, achieves $O(d^2 \log^2 n)$ measurements[4].

Approximate reconstruction of sparse vectors up to a small number of false positives (that is one focus of this work) has been studied as a major ingredient of trivial two-stage schemes [17,7,25,26,27,8]. In particular, a generalization of superimposed codes, known as *selectors*, was introduced in [26] which, roughly speaking, allows for identification of the sparse vector up to a prescribed number of false positives. They gave a non-constructive result showing that there are such (non-adaptive) schemes that keep the number of false positives at $O(d)$ using $O(d \log(n/d))$ measurements, matching the optimal "counting bound". A probabilistic construction of asymptotically optimal selectors (resp., a related notion of *resolvable matrices*) is given in [8] (resp., [27]), and [28,29] give slightly sub-optimal "explicit" constructions based on certain expander graphs obtained from dispersers[5].

To give a concise comparison of the present work with those listed above, we mention some of the qualities of the group testing schemes that we will aim to attain: (1) low number of measurements, (2) arbitrarily good degree of approximation, (3) maximum possible noise tolerance, (4) efficient, deterministic construction: As typically the sparsity d is very small compared to n, a measurement matrix must be ideally *fully explicitly constructible* in the sense that each entry of the matrix should be computable in deterministic time $\mathsf{poly}(d, \log n)$ (e.g., while the constructions in [26,27,8,28,29] are all polynomial-time computable in n, they are not fully explicit in this sense). (5) fully efficient reconstruction algorithm: For a similar reason, the length of the observation vector is typically far smaller than n; thus, it is desirable to have a reconstruction algorithm that identifies the support of the sparse vector in time polynomial in the number of measurements (which might be exponentially smaller than n). While the works that we mentioned focus on few of the criteria listed above (e.g., none of the above-mentioned schemes for approximate group testing are equipped with a fully efficient reconstruction algorithm), our approach can potentially attain *all*

[4] Interestingly, this classical construction can be regarded as a special instantiation of our framework where a "bounded degree univariate polynomial" is used in place of the underlying randomness condenser. However, the analysis and the properties of the resulting group testing schemes substantially differ for the two cases, and in particular, the MDS-based construction owes its properties essentially to the large distance of the underlying code. In Appendix B, we will elaborate in more detail on this correspondence as well as a connection with the *bit-probe* model in data structures.

[5] The notion of selectors is useful in a noiseless setting. However, as remarked in [8], it can be naturally extended to include a "noise" parameter, and the probabilistic constructions of selectors can be naturally extended to this case. Nonetheless, this generalization does not distinguish between false positives and negatives and the explicit constructions of selectors [28,29] cannot be used in a (highly) noisy setting.

at the same time. As we will see later, using the best known constructions of condensers we will have to settle to sub-optimal results in one or more of the aspects above. Nevertheless, the fact that any improvement in the construction of condensers would readily translate to improved group testing schemes (and also the rapid growth of derandomization theory) justifies the significance of the construction given in this work.

2 Preliminaries

For non-negative integers e_0 and e_1, we say that an ordered pair of binary vectors (x, y), each in $\mathbb{F}_2^n$, are (e_0, e_1)-close (or x is (e_0, e_1)-close to y) if y can be obtained from x by flipping at most e_0 bits from 0 to 1 and at most e_1 bits from 1 to 0. Hence, such x and y will be $(e_0 + e_1)$-close in Hamming-distance. Further, (x, y) are called (e_0, e_1)-far if they are not (e_0, e_1)-close. Note that if x and y are seen as characteristic vectors of subsets X and Y of $[n]$, respectively[6], they are $(|Y \setminus X|, |X \setminus Y|)$-close. Furthermore, (x, y) are (e_0, e_1)-close iff (y, x) are (e_1, e_0)-close. A group of m non-adaptive measurements for binary vectors of length n can be seen as an $m \times n$ matrix (that we call the *measurement matrix*) whose (i, j)th entry is 1 iff the jth coordinate of the vector is present in the disjunction defining the ith measurement. For a measurement matrix A, we denote by $A[x]$ the outcome of the measurements defined by A on a binary vector x, that is, the bitwise OR of those columns of A chosen by the support of x. As motivated by our negative results, for the specific setting of the group testing problem that we are considering in this work, it is necessary to give an *asymmetric* treatment that distinguishes between inaccuracies due to false positives and false negatives. Thus, we will work with a notion of error-tolerating measurement matrices that directly and conveniently captures this requirement, as given below:

Definition 1. Let $m, n, d, e_0, e_1, e_0', e_1'$ be integers. An $m \times n$ measurement matrix A is called (e_0, e_1, e_0', e_1')-correcting for d-sparse vectors if, for every $y \in \mathbb{F}_2^m$ there exists $z \in \mathbb{F}_2^n$ (called a *valid decoding of* y) such that for every $x \in \mathbb{F}_2^n$, whenever (x, z) are (e_0', e_1')-far, $(A[x], y)$ are (e_0, e_1)-far. The matrix A is called *fully explicit* if each entry of the matrix can be computed in time $\mathsf{poly}(\log n)$.

Intuitively, the definition states that two measurements are allowed to be confused only if they are produced from close vectors. In particular, an (e_0, e_1, e_0', e_1')-correcting matrix gives a group testing scheme that reconstructs the sparse vector up to e_0' false positives and e_1' false negatives even in the presence of e_0 false positives and e_1 false negatives in the measurement outcome. Under this notation, unique decoding would be possible using an $(e_0, e_1, 0, 0)$-correcting matrix if the amount of measurement errors is bounded by at most e_0 false positives and e_1 false negatives. However, when $e_0' + e_1'$ is positive, decoding may require a bounded amount of ambiguity, namely, up to e_0' false positives and e_1' false negatives in the decoded sequence. In the combinatorics literature, the special case of

[6] We use the shorthand $[n]$ for the set $\{1, 2, \ldots, n\}$.

$(0,0,0,0)$-correcting matrices is known as *d-superimposed codes* or *d-separable matrices* and is closely related to the notions of *d-cover-free families* and *d-disjunct* matrices (cf. [15] for precise definitions). Also, $(0,0,e'_0,0)$-correcting matrices are related to the notion of *selectors* in [26] and *resolvable matrices* in [27].

The *min-entropy* of a distribution $\mathcal{X}$ with finite support S is given by $H_\infty(\mathcal{X}) := \min_{x\in S}\{-\log \Pr_{\mathcal{X}}(x)\}$, where $\Pr_{\mathcal{X}}(x)$ is the probability that $\mathcal{X}$ assigns to x. The *statistical distance* of two distributions $\mathcal{X}$ and $\mathcal{Y}$ defined on the same finite space S is given by $\frac{1}{2}\sum_{s\in S}|\Pr_{\mathcal{X}}(s) - \Pr_{\mathcal{Y}}(s)|$, which is half the ℓ_1 distance of the two distributions when regarded as vectors of probabilities over S. Two distributions $\mathcal{X}$ and $\mathcal{Y}$ are said to be ϵ-close if their statistical distance is at most ϵ. We will use the shorthand $\mathcal{U}_n$ for the uniform distribution on $\mathbb{F}_2^n$, and $X \sim \mathcal{X}$ for a random variable X drawn from a distribution $\mathcal{X}$. A function $C\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ is a *strong* $k \to_\epsilon k'$ *condenser* if for every distribution $\mathcal{X}$ on $\mathbb{F}_2^n$ with min-entropy at least k, random variable $X \sim \mathcal{X}$ and a *seed* $Y \sim \mathcal{U}_t$, the distribution of $(Y, C(X,Y))$ is ϵ-close to some distribution $(\mathcal{U}_t, \mathcal{Z})$ with min-entropy at least $t + k'$. The parameters ϵ, $k - k'$, and $\ell - k'$ are called the *error*, the *entropy loss* and the *overhead* of the condenser, respectively. A condenser with zero entropy loss is called *lossless*, and a condenser with zero overhead is called a *strong* (k,ϵ)*-extractor*. A condenser is *explicit* if it is polynomial-time computable.

3 Negative Results

In coding theory, it is possible to construct codes that can tolerate up to a constant fraction of adversarially chosen errors and still guarantee unique decoding. Hence it is natural to wonder whether a similar possibility exists in group testing, namely, whether there is a measurement matrix that is robust against a constant fraction of adversarial errors and still recovers the measured vector exactly. Below we show that this is not possible[7]:

Lemma 2. *Suppose that an $m \times n$ measurement matrix A is (e_0, e_1, e'_0, e'_1)-correcting for d-sparse vectors. Then $(\max\{e_0, e_1\} + 1)/(e'_0 + e'_1 + 1) \le m/d$.* □

The above lemma (proved in in Appendix C.1) gives a trade-off between the tolerable error in the measurements versus the reconstruction error. In particular, for unique decoding to be possible one can only guarantee resiliency against up to $O(1/d)$ fraction of errors in the measurement. On the other hand, tolerance against a constant fraction of errors would make an ambiguity of order $\Omega(d)$ in the decoding inevitable. Another trade-off is given by the following lemma (proved in Appendix C.2):

[7] We remark that the negative results in this section hold for both adaptive and non-adaptive measurements.

Lemma 3. *Suppose that an $m \times n$ measurement matrix A is (e_0, e_1, e'_0, e'_1)-correcting for d-sparse vectors. Then for every $\epsilon > 0$, either $e_1 < \frac{(e'_1+1)m}{\epsilon d}$ or $e'_0 \geq \frac{(1-\epsilon)(n-d+1)}{(e'_1+1)^2}$.* □

As mentioned in the introduction, it is an important matter for applications to bring down the amount of false negatives in the reconstruction as much as possible, and ideally to zero. The lemma above shows that if one is willing to keep the number e'_1 of false negatives in the reconstruction at the zero level (or bounded by a constant), only an up to $O(1/d)$ fraction of false negatives in the measurements can be tolerated (regardless of the number of measurements), unless the number e'_0 of false positives in the reconstruction grows to an enormous amount (namely, $\Omega(n)$ when $n - d = \Omega(n)$) which is certainly undesirable.

As shown in [21], exact reconstruction of d-sparse vectors of length n, even in a noise-free setting, requires at least $\Omega(d^2 \log n / \log d)$ non-adaptive measurements. However, it turns out that there is no such restriction when an approximate reconstruction is sought for, except for the following bound which can be shown using simple counting and holds for adaptive noiseless schemes as well (proof in Appendix C.3):

Lemma 4. *Let A be an $m \times n$ measurement matrix that is $(0, 0, e'_0, e'_1)$-correcting for d-sparse vectors. Then $m \geq d \log(n/d) - d - e'_0 - O(e'_1 \log((n - d - e'_0)/e'_1))$, where the last term is defined to be zero for $e'_1 = 0$.* □

This is similar in spirit to the lower bound obtained in [26] for the size of selectors. According to the lemma, even in the noiseless scenario, any reconstruction method that returns an approximation of the sparse vector up to $e'_0 = O(d)$ false positives and without false negatives will require $\Omega(d \log(n/d))$ measurements. As we will show in the next section, an upper bound of $O(d \log n)$ is in fact attainable even in a highly noisy setting using only non-adaptive measurements. This in particular implies an asymptotically optimal trivial two-stage group testing scheme.

4 A Noise-Resilient Construction

In this section we give our general construction and design measurement matrices for testing D-sparse vectors[8] in $\mathbb{F}_2^N$. The matrices can be seen as adjacency matrices of certain unbalanced bipartite graphs constructed from good randomness condensers or extractors. The main technique that we use to show the desired properties is the *list-decoding view* of randomness condensers, extractors, and expanders, developed over the recent years starting from the work of Ta-Shma and Zuckerman on *extractor codes* [30]. We start by introducing the terms that we will use in this construction and the analysis.

[8] In this section we find it more convenient to use capital letters $D, N, \ldots$ instead of $d, n, \ldots$ that we have so far used and keep the small letters for their base-2 logarithms.

Definition 5. (mixtures, agreement, and agreement list) Let Σ be a finite set. A *mixture* over Σ^n is an n-tuple $S := (S_1, \ldots, S_n)$ such that every S_i, $i \in [n]$, is a nonempty subset of Σ. The *agreement* of $w := (w_1, \ldots w_n) \in \Sigma^n$ with S, denoted by $\mathsf{Agr}(w, S)$, is the quantity $\frac{1}{n}|\{i \in [n] \colon w_i \in S_i\}|$. Moreover, we define the quantity $\mathsf{wgt}(S) := \sum_{i \in [n]} |S_i|$ and $\rho(S) := \mathsf{wgt}(S)/(n|\Sigma|)$, where the latter is the expected agreement of a random vector with S. For a code $\mathcal{C} \subseteq \Sigma^n$ and $\alpha \in (0, 1]$, the *α-agreement list* of $\mathcal{C}$ with respect to S, denoted by $\mathsf{LIST}_{\mathcal{C}}(S, \alpha)$, is the set[9] $\mathsf{LIST}_{\mathcal{C}}(S, \alpha) := \{c \in \mathcal{C} \colon \mathsf{Agr}(c, S) > \alpha\}$.

Definition 6. (induced code) Let $f \colon \Gamma \times \Omega \to \Sigma$ be a function mapping a finite set $\Gamma \times \Omega$ to a finite set Σ. For $x \in \Gamma$, we use the shorthand $f(x)$ to denote the vector $y := (y_i)_{i \in \Omega}$, $y_i := f(x, i)$, whose coordinates are indexed by the elements of Ω in a fixed order. The *code induced by f*, denoted by $\mathcal{C}(f)$ is the set $\{f(x) \colon x \in \Gamma\}$. The induced code has a natural encoding function given by $x \mapsto f(x)$.

Definition 7. (codeword graph) Let $\mathcal{C} \subseteq \Sigma^n$, $|\Sigma| = q$, be a q-ary code. The *codeword graph* of $\mathcal{C}$ is a bipartite graph with left vertex set $\mathcal{C}$ and right vertex set $n \times \Sigma$, such that for every $x = (x_1, \ldots, x_n) \in \mathcal{C}$, there is an edge between x on the left and $(1, x_1), \ldots, (n, x_n)$ on the right. The *adjacency matrix* of the codeword graph is an $n|\Sigma| \times |\mathcal{C}|$ binary matrix whose (i, j)th entry is 1 iff there is an edge between the ith right vertex and the jth left vertex.

The following is a straightforward generalization of the result in [30] that is also shown in [31] (we have included a proof in Appendix C.4):

Theorem 8. *Let $f \colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ be a strong $k \to_\epsilon k'$ condenser, and $\mathcal{C} \subseteq \Sigma^{2^t}$ be its induced code, where $\Sigma := \mathbb{F}_2^\ell$. Then for any mixture S over Σ^{2^t} we have $|\mathsf{LIST}_{\mathcal{C}}(S, \rho(S)2^{\ell - k'} + \epsilon)| < 2^k$.* □

Now using the above tools, we are ready to describe our construction of error-tolerant measurement matrices. We first state a general result without specifying the parameters of the condenser, and then instantiate the construction with various choices of the condenser, resulting in matrices with different properties.

Theorem 9. *Let $f \colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ be a strong $k \to_\epsilon k'$ condenser, and $\mathcal{C}$ be its induced code, and define the capital shorthands $K := 2^k$, $K' := 2^{k'}$, $L := 2^\ell$, $N := 2^n$, $T := 2^t$. Suppose that the parameters $p, \nu, \gamma > 0$ are chosen such that $(p + \gamma)L/K' + \nu/\gamma < 1 - \epsilon$, and $D := \gamma L$. Then the adjacency matrix of the codeword graph of $\mathcal{C}$ (which has $M := TL$ rows and N columns) is a $(pM, (\nu/D)M, K - D, 0)$-correcting measurement matrix for D-sparse vectors. Moreover, it allows for a reconstruction algorithm with running time $O(MN)$.*

A proof of the theorem is given in Appendix C.5, which uses Theorem 8 as an essential tool. Here we recall a description of the reconstruction algorithm from the proof: Let $\hat{y} \in \mathbb{F}_2^{TL}$ be the observation outcome, and m_{ij} denote the (i, j)th

[9] When $\alpha = 1$, we consider codewords with full agreement with the mixture.

entry of the measurement matrix. Then the reconstruction algorithm outputs a vector $\hat{x} \in \mathbb{F}_2^N$ that satisfies $(\forall i \in [N])\ \hat{x}_i = 1$ iff $|\{j \in [TL] \colon \hat{y}_j = m_{ji} = 1\}| \geq T(1 - \nu/\gamma)$.

Remark 1. Extractor codes that we use in Theorem 9 are instances of *soft-decision decodable* codes[10] that provide high list-decodability in "extremely noisy" scenarios. In fact it is not hard to see that good extractors or condensers are required for our construction to carry through, as Theorem 8 can be shown to hold in the reverse direction as well. However, for designing measurement matrices for the noiseless (or low-noise) case, it is possible to resort to the slightly weaker notion of *list recoverable codes*. This is discussed in more detail in Appendix A.

Tolerance on Incorrect Estimates. The result given by Theorem 9 requires at least an overestimate on the number of defectives (i.e., the sparsity level D that is controlled by the parameter γ) and the level of measurement noise (given by the parameters p and ν). However, if in the actual experiment the estimates turn out to be incorrect but the trade-off on γ, p, ν required by the theorem remains valid (e.g., if the number of defectives turns out higher but the fraction of measurement errors lower than expected) we can still guarantee a reliable reconstruction. In case there is no reliable estimate on D available, one can use a number of trial and error rounds by starting from an initial guess of $D = 1$ and doubling the guess in each round, until at some point the Hamming weight of the reconstruction does not exceed the amount K guaranteed by the theorem. For all our instantiations that follow, the total number of measurements required by this process remains in the same order as if we knew the actual value of D.

Instantiations

Now we instantiate the general result given by Theorem 9 with various choices of the underlying condenser and compare the obtained parameters. First, we consider two extreme cases, namely, a non-explicit optimal condenser with zero overhead (i.e., extractor) and then a non-explicit optimal condenser with zero loss (i.e., lossless condenser) and then consider how known explicit constructions can approach the obtained bounds. We remark that the *sampling rate*, as defined in [27], of these instantiations (i.e., the maximum number of tests any individual is included in) is $O(1/D)$ fraction of the number of tests.

Applying Optimal Extractors. Radhakrishan and Ta-Shma showed that non-constructively, for every k, n, ϵ, there is a strong (k, ϵ)-extractor with seed length $t = \log(n - k) + 2\log(1/\epsilon) + O(1)$ and output length $\ell = k - 2\log(1/\epsilon) - O(1)$, which is the best one can hope for [32]. In particular, they show that a random function achieves these parameters with probability $1 - o(1)$. Plugging this result in Theorem 9, we obtain a non-explicit measurement matrix from a simple, randomized construction that achieves the desired trade-off with high probability (see Appendix C.6 for the proof details):

[10] To be precise, here we are dealing with a special case of soft-decision decoding with binary weights.

Corollary 10. *For every choice of constants $p \in [0,1)$ and $\nu \in [0,\nu_0)$, $\nu_0 := (\sqrt{5-4p}-1)^3/8$, and positive integers D and $N \geq D$, there is an $M \times N$ measurement matrix, where $M = O(D \log N)$, that is $(pM, (\nu/D)M, O(D), 0)$-correcting for D-sparse vectors of length N and allows for a reconstruction algorithm with running time $O(MN)$.* □

This instantiation, in particular, reproduces a result on randomized construction of approximate group testing schemes with optimal number of measurements in [8], but with stringent conditions on the noise tolerance of the scheme.

Applying Optimal Lossless Condensers. The probabilistic construction of Radhakrishan and Ta-Shma can be extended to the case of lossless condensers and one can show that a random function is with high probability a strong $k \to_\epsilon k$ condenser with seed length $t = \log n + \log(1/\epsilon) + O(1)$ and output length $\ell = k + \log(1/\epsilon) + O(1)$ [33]. This combined with Theorem 9 gives the following corollary (proof in Appendix C.7):

Corollary 11. *For positive integers $N \geq D$ and every constant $\delta > 0$ there is an $M \times N$ measurement matrix, where $M = O(D \log N)$, that is $(\Omega(M), \Omega(1/D)M, \delta D, 0)$-correcting for D-sparse vectors of length N and allows for a reconstruction algorithm with running time $O(MN)$.* □

Both results obtained in Corollaries 10 and 11 almost match the lower bound of Lemma 4 for the number of measurements. However, we note the following distinction between the two results: Instantiating the general construction of Theorem 9 with an extractor gives us a sharp control over the fraction of tolerable errors, and in particular, we can obtain a measurement matrix that is robust against *any* constant fraction (bounded from 1) of false positives. However, the number of false positives in the reconstruction will be bounded by some constant fraction of the sparsity of the vector that cannot be made arbitrarily close to zero. On the other hand, a lossless condenser enables us to bring down the number of false positives in the reconstruction to an arbitrarily small fraction of D (which is, in light of Lemma 2, the best we can hope for), but on the other hand, does not give as good a control on the fraction of tolerable errors as in the extractor case, though we still obtain resilience against the same order of errors.

Applying the Guruswami-Umans-Vadhan's Extractor. While Corollaries 10 and 11 give probabilistic constructions of noise-resilient measurement matrices, certain applications require a fully explicit matrix that is guaranteed to work. To that end, we need to instantiate Theorem 9 with an explicit condenser. First, we use a nearly-optimal explicit extractor due to Guruswami, Umans and Vadhan, summarized in the following theorem:

Theorem 12. *[31] For all positive integers $n \geq k$ and all $\epsilon > 0$, there is an explicit strong (k,ϵ)-extractor $\mathsf{Ext}\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ with $\ell = k - 2\log(1/\epsilon) - O(1)$ and $t = \log n + O(\log k \cdot \log(k/\epsilon))$.* □

Applying this result in Theorem 9 we obtain a similar trade-off as in Corollary 10, except for a higher number of measurements which would be bounded by $O(2^{O(\log^2 \log D)} D \log N) = O(D^{1+o(1)} \log N)$.

Applying the Zig-Zag Lossless Condenser. In [33] an explicit lossless condenser with optimal output length is constructed. In particular they show the following:

Theorem 13. *[33] For every $k \leq n \in \mathbb{N}$, $\epsilon > 0$ there is an explicit $k \to_\epsilon k$ condenser*[11] *with seed length $O(\log^3(n/\epsilon))$ and output length $k + \log(1/\epsilon) + O(1)$.*

Combined with Theorem 9, we obtain a similar result as in Corollary 11, except that the number of measurements would be $D2^{\log^3(\log N)} = D \cdot \mathsf{quasipoly}(\log N)$.

Measurements Allowing Sublinear Time Reconstruction. The naive reconstruction algorithm given by Theorem 9 works efficiently in linear time in the size of the measurement matrix. However, as mentioned in the introduction, for very sparse vectors (i.e., $D \ll N$) it might be of practical importance to have a reconstruction algorithm that runs in *sublinear* time in N, the length of the vector, and ideally, polynomial in the number of measurements, which is merely $\mathsf{poly}(\log N, D)$ if the number of measurements is optimal.

As shown in [30], if the code $\mathcal{C}$ in Theorem 8 is obtained from a strong extractor constructed from a *black-box pseudorandom generator (PRG)*, it is possible to compute the agreement list (which is guaranteed by the theorem to be small) more efficiently than a simple exhaustive search over all possible codewords. In particular, in this case they show that $\mathsf{LIST}_{\mathcal{C}}(S, \rho(S) + \epsilon)$ can be computed in time $\mathsf{poly}(2^t, 2^\ell, 2^k, 1/\epsilon)$, which can be much smaller than 2^n. On the other hand, observe that the main computational task done by the reconstruction algorithm in Theorem 9 is in fact computing a suitable agreement list for the induced code of the underlying condenser.

Currently two constructions of extractors from black-box PRGs are known: Trevisan's extractor [34] (as well as its improvement in [35]) and Shaltiel-Umans' extractor [36]. However, the latter can only extract a sub-constant fraction of the min-entropy and is not suitable for our needs, albeit it requires a considerably shorter seed than Trevisan's extractor. Thus, here we only consider an improvement of Trevisan's extractor given by Raz *et al.*, quoted below.

Theorem 14. *[35] For every $n, k, \ell \in \mathbb{N}$, $(\ell \leq k \leq n)$ and $\epsilon > 0$, there is an explicit strong (k, ϵ)-extractor $\mathsf{Tre}\colon \mathbb{F}_2^n \times \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ with $t = O(\log^2(n/\epsilon) \cdot \log(1/\alpha))$, where $\alpha := k/(\ell - 1) - 1$ must be less than $1/2$.* □

Using this result in Theorem 9, we obtain a measurement matrix for which the reconstruction is possible in polynomial time in the number of measurements; however, as the seed length required by this extractor is larger than Theorem 12, we will now require a higher number of measurements than before. Specifically, we obtain the same parameters as in Corollary 10 using Trevisan's extractor except for the number of measurements, $M = O(D2^{\log^3 \log N}) = D \cdot \mathsf{quasipoly}(\log N)$.

[11] Though not explicitly mentioned in [33], these condensers can be considered to be strong.

Furthermore, Guruswami *et al.* [31] construct lossless (and lossy) condensers that are not known to correspond to black-box PRGs but however allow efficient list-recovery. In particular they show the following:

Theorem 15. *[31] For all constants $\alpha \in (0,1)$ and every $k \leq n \in \mathbb{N}$, $\epsilon > 0$ there is an explicit strong $k \to_\epsilon k$ condenser with seed length $t = (1+1/\alpha)\log(nk/\epsilon) + O(1)$ and output length $\ell = d + (1+\alpha)k$. Moreover, the condenser has efficient list recovery.* □

The code induced by the condenser given by this theorem is precisely a Parvaresh-Vardy code [37] and thus, the efficient list recovery is merely the list-decoding algorithm for this code. Combined with Theorem 9 we can show that codeword graphs of Parvaresh-Vardy codes correspond to good measurement matrices that allow sublinear time recovery, but with incomparable parameters to what we obtained from Trevisan's extractor (the proof is similar to Corollary 11):

Corollary 16. *For positive integers $N \geq D$ and any constants $\delta, \alpha > 0$ there is an $M \times N$ measurement matrix, where $M = O(D^{3+\alpha+2/\alpha}(\log N)^{2+2/\alpha})$, that is $(\Omega(e), \Omega(e/D), \delta D, 0)$-correcting for D-sparse vectors of length N, where $e := (\log N)^{1+1/\alpha} D^{2+1/\alpha}$. Moreover, the matrix allows for a reconstruction algorithm with running time $\mathsf{poly}(M)$.* □

We remark that we could also use a lossless condenser due to Ta-Shma *et al.* [38] which is based on Trevisan's extractor and also allows efficient list recovery, but it achieves inferior parameters compared to Corollary 16.

Future Work

For the purpose of this exposition, we have focused on the asymptotic trade-offs and on several occasions have neglected certain details such as the hidden constants in the $O(\cdot)$ notation that become important for practical purposes. We defer the task of estimating and optimizing for these parameters as well as obtaining experimental results to the subsequent work. Moreover, an interesting theoretical question to ask is whether our reduction from group testing schemes to construction of condensers holds in the reverse direction as well; namely, whether one can obtain a a good condenser from *any* highly noise-resilient group testing scheme.

Acknowledgment

The author is thankful to Amin Shokrollahi for introducing him to the group testing problem and his comments on an earlier draft of this paper, and to Venkatesan Guruswami for several illuminating discussions that led to considerable improvement of the results presented in this work.

References

1. Dorfman, R.: The detection of defective members of large populations. Annals of Mathematical Statistics **14** (1943) 436–440
2. Sobel, M., Groll, P.: Group-testing to eliminate efficiently all defectives in a binomial sample. Bell Systems Technical Journal **38** (1959) 1179–1252
3. Pevzner, P., Lipshutz, R.: Towards DNA sequencing chips. In: Proceedings of MFCS. Volume 841 of Lecture Notes in Computer Science. (1994) 143–158
4. Ngo, H., Du, D.: A survey on combinatorial group testing algorithms with applications to DNA library screening. DIMACS Series on Discrete Math. and Theoretical Computer Science **55** (2000) 171–182
5. Schliep, A., Torney, D., Rahmann, S.: Group testing with DNA chips: Generating designs and decoding experiments. In: Proc. Computational Syst. Bioinformatics. (2003)
6. Macula, A.: Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. Annals of Combinatorics **3**(1) (1999) 61–69
7. Macula, A.: Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. J. Comb. Optim. **2** (1999) 385–397
8. Cheng, Y., Du, D.Z.: New constructions of one- and two-stage pooling designs. Journal of Computational Biology **15**(2) (2008) 195–205
9. Wolf, J.: Born-again group testing: multiaccess communications. IEEE Transactions on Information Theory **31** (1985) 185–191
10. Hong, E., Ladner, R.: Group testing for image compression. In: Data Compression Conference. (2000) 3–12
11. Clifford, R., Efremenko, K., Porat, E., Rothschild, A.: k-mismatch with don't cares. In: Proceedings of ESA. Volume 4698 of LNCS. (2007) 151–162
12. Cormode, G., Muthukrishnan, S.: What's hot and what's not: tracking most frequent items dynamically. ACM Trans. on Database Syst. **30**(1) (2005) 249–278
13. Blass, A., Gurevich, Y.: Pairwise testing. Bulletin of the EATCS **78** (2002) 100–132
14. Cormode, G., Muthukrishnan, S.: Combinatorial algorithms for compressed sensing. In: Proceedings of Information Sciences and Systems. (2006) 198–201
15. Du, D.Z., Hwang, F.: Combinatorial Group Testing and its Applications. Second edn. World Scientific (2000)
16. Du, D.Z., Hwang, F.: Pooling Designs and Nonadaptive Group Testing. World Scientific (2006)
17. Knill, E.: Lower bounds for identifying subset members with subset queries. In: Proceedings of SODA. (1995) 369–377
18. Dyachkov, A., Rykov, V.: A survey of superimposed code theory. Problems of Control and Information Theory **12**(4) (1983) 229–242
19. Knill, E., Bruno, W.J., Torney, D.C.: Non-adaptive group testing in the presence of errors. Discrete Appl. Math. **88**(1–3) (1998) 261–290
20. Macula, A.: Error correcting nonadaptive group testing with d^e-disjunct matrices. Discrete Applied Mathematics **80**(2,3) (1997) 217–222
21. D'yachkov, A., Rykov, V.: Bounds of the length of disjunct codes. Problems of Control and Information Theory **11** (1982) 7–13
22. Ruszinkó: On the upper bound of the size of the r-cover-free families. J. Combin. Thy., series A **66** (1994) 302–310
23. Füredi, Z.: On r-cover-free families. Journal of Combinatorial Theory, Series A **73** (1996) 172–173

24. Kautz, W., Singleton, R.: Nonrandom binary superimposed codes. IEEE Transactions on Information Theory **10** (1964) 363–377
25. Berger, T., Mandell, J., Subrahmanya, P.: Maximally efficient two-stage group testing. Biometrics **56** (2000) 833–840
26. De Bonis, A., Gasieniec, L., Vaccaro, U.: Optimal two-stage algorithms for group testing problems. SIAM Journal on Computing **34**(5) (2005) 1253–1270
27. Eppstein, D., Goodrich, M., Hirschberg, D.: Improved combinatorial group testing algorithms for real-world problem sizes. SIAM Journal on Computing **36**(5) (2007) 1360–1375
28. Indyk, P.: Explicit constructions of selectors with applications. In: Proceedings of SODA. (2002)
29. Chlebus, B., Kowalski, D.: Almost optimal explicit selectors. In: Proceedings of FCT. Volume 3623 of Lecture Notes in Computer Science. (2005) 270–280
30. Ta-Shma, A., Zuckerman, D.: Extractor codes. IEEE Transactions on Information Theory **50**(12) (2004) 3015–3025
31. Guruswami, V., Umans, C., Vadhan, S.: Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. In: Proc. of the 22nd IEEE CCC. (2007)
32. Radhakrishan, J., Ta-Shma, A.: Tight bounds for depth-two superconcentrators. In: Proceedings of the 38th FOCS. (1997) 585–594
33. Capalbo, M., Reingold, O., Vadhan, S., Wigderson, A.: Randomness conductors and constant-degree expansion beyond the degree/2 barrier. In: Proceedings of the 34th STOC. (2002) 659–668
34. Trevisan, L.: Extractors and pseudorandom generators. Journal of the ACM **48**(4) (2001) 860–879
35. Raz, R., Reingold, O., Vadhan, S.: Extracting all the randomness and reducing the error in Trevisan's extractor. JCSS **65**(1) (2002) 97–128
36. Shaltiel, R., Umans, C.: Simple extractors for all min-entropies and a new pseudorandom generator. Journal of the ACM **52**(2) (2005) 172–216
37. Parvaresh, F., Vardy, A.: Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In: Proceedings of the 46th FOCS. (2005) 285–294
38. Ta-Shma, A., Umans, C., Zuckerman, D.: Lossless condensers, unbalanced expanders, and extractors. In: Proceedings of the 33th STOC. (2001) 143–152
39. Guruswami, V.: List Decoding of Error-Correcting Codes. PhD thesis, Massachusetts Institute of Technology (2001)
40. Guruswami, V., Rudra, A.: Concatenated codes can achieve list decoding capacity. In: Proceedings of SODA. (2008)
41. Buhrman, H., Miltersen, P., Radhakrishnan, J., Venkatesh, S.: Are bitvectors optimal? SIAM Journal on Computing **31**(6) (2002) 1723–1744
42. Ta-Shma, A.: Storing information with extractors. Information Processing Letters **83**(5) (2002) 267–274
43. Buresh-Oppenheim, J., Kabanets, V., Santhanam, R.: Uniform hardness amplification in NP via monotone codes. *ECCC* Technical Report TR06-154. (2006)
44. Guruswami, V., Gopalan, P.: Hardness amplification within NP against deterministic algorithms. In: Proceedings of the 23rd IEEE CCC. (2008) 19–30

A Connection with List-Recoverability

A pointed out in Remark 1, measurement matrices that approximate sparse vectors using a small number of noiseless measurements can be constructed from

list recoverable codes. Formally, a code $\mathcal{C}$ of block length n over an alphabet Σ is called (α, D, L)-*list recoverable* if for every mixture S over Σ^n consisting of sets of size at most D each, we have $|\mathsf{LIST}_{\mathcal{C}}(S, \alpha)| \leq L$. A simple argument similar to Theorem 9 shows that the adjacency matrix of the codeword graph of such a code with rate R gives a $(\log N)|\Sigma|/R \times N$ measurement matrix[12] for D-sparse vectors in the noiseless case with at most $L - D$ false positives in the reconstruction. Ideally, a list-recoverable code with $\alpha = 1$, alphabet size $O(D)$, positive constant rate, and list size $L = O(D)$ would give an $O(D \log N) \times N$ matrix for D-sparse vectors, which is almost optimal (furthermore, the recovery would be possible in sublinear time if $\mathcal{C}$ is equipped with efficient list recovery). However, no explicit construction of such a code is so far known.

Two natural choices of list-recoverable codes are Reed-Solomon and Algebraic-Geometric codes, which in fact provide soft-decision decoding with short list size (cf. [39]). However, while the list size is polynomially bounded by n and D, it can be much larger than $O(D)$ that we need for our application even if the rate is polynomially small in D. On the other hand, it is shown in [40] that *folded Reed-Solomon Codes* are list-recoverable with constant rate, but again they suffer from large alphabet and list size[13]. We also point out a construction of (α, D, D) list-recoverable codes (allowing list recovery in time $O(nD)$) in [40] with rate polynomially small but alphabet size exponentially large in D, from which they obtain superimposed codes.

B Connection with the Bit-Probe Model and Designs

An important problem in data structures is the static set membership problem in bit-probe model, which is the following: Given a set S of at most d elements from a universe of size n, store the set as a string of length m such that any query of the type "is x in S?" can be reliably answered by reading few bits of the encoding. The query algorithm might be probabilistic, and be allowed to err with a small one or two-sided error. Information theoretically, it is easy to see that $m = \Omega(d \log(n/d))$ regardless of the bit-probe complexity and even if a small constant error is allowed.

Remarkably, it was shown in [41] that the lower bound on m can be (non-explicitly) achieved using only one bit-probe. Moreover, a part of their work shows that any one-probe scheme with negative one-sided error ϵ (where the scheme only errs in case $x \notin S$) gives a $\lfloor d/\epsilon \rfloor$-superimposed code (and hence, requires $m = \Omega(d^2 \log n)$ by [21]). It follows that from any such scheme one can obtain a measurement matrix for exact reconstruction of sparse vectors, which, by Lemma 2, cannot provide high resiliency against noise. The converse direction, i.e., using superimposed codes to design bit-probe schemes does not

[12] For codes over large alphabets, the factor $|\Sigma|$ in the number of rows can be improved using *concatenation* with a suitable *inner* measurement matrix.

[13] As shown in [31], folded Reed-Solomon codes can be used to construct lossless condensers, which eliminates the list size problem. However, they give inferior parameters compared to Parvaresh-Vardy codes used in Corollary 16.

necessarily hold unless the error is allowed to be very close to 1. However, in [41] *combinatorial designs*[14] based on low-degree polynomials are used to construct one bit-probe schemes with $m = O(d^2 \log^2 n)$ and small one-sided error.

On the other hand, Kautz and Singleton [24] observed that the encoding of a combinatorial design as a binary matrix corresponds to a superimposed code (which is in fact slightly error-resilient). Moreover, they used Reed-Solomon codes to construct a design, which in particular gives a d-superimposed code. This is in fact the same design that is used in [41], and in our terminology, can be regarded as the adjacency matrix of the codeword graph of a Reed-Solomon code. It is interesting to observe the intimate similarity between our framework given by Theorem 9 and classical constructions of superimposed codes. However, some key differences are worth mentioning. Indeed, both constructions are based on codeword graphs of error-correcting codes. However, classical superimposed codes owe their properties to the large distance of the underlying code. On the other hand, our construction uses extractor and condenser codes and does not give a superimposed code simply because of the substantially low number of measurements. However, as shown in Theorem 9, they are good enough for a slight relaxation of the notion of superimposed codes because of their soft-decision list decodability properties, which additionally enables us to attain high noise resilience and a considerably smaller number of measurements.

Interestingly, Buhrman *et al.* [41] use randomly chosen bipartite graphs to construct storage schemes with two-sided error requiring nearly optimal space $O(d \log n)$, and Ta-Shma [42] later shows that expander graphs from lossless condensers would be sufficient for this purpose. However, unlike schemes with negative one-sided error, these schemes use encoders that cannot be implemented by the OR function and thus do not translate to group-testing schemes.

C Omitted Proofs

C.1 Proof of Lemma 2

We use similar arguments as those used in [43,44] in the context of black-box hardness amplification in NP: Define a partial ordering $\prec$ between binary vectors using bit-wise comparisons (with $0 < 1$). Let $t := d/(e'_0 + e'_1 + 1)$ be an integer[15], and consider any monotonically increasing sequence of vectors $x_0 \prec \cdots \prec x_t$ in $\mathbb{F}_2^n$ where x_i has weight $i(e'_0 + e'_1 + 1)$. Thus, x_0 and x_t will have weights zero and d, respectively. Note that we must also have $A[x_0] \prec \cdots \prec A[x_t]$ due to monotonicity of the OR function.

A fact that is directly deduced from Definition 1 is that, for every $x, x' \in \mathbb{F}_2^n$, if $(A[x], A[x'])$ are (e_0, e_1)-close, then x and x' must be $(e'_0+e'_1, e'_0+e'_1)$-close. This

[14] A design is a collection of subsets of a universe, each of the same size, such that the pairwise intersection of any two subset is upper bounded by a prespecified parameter.

[15] For the sake of simplicity in this presentation we ignore the fact that certain fractions might in general give non-integer values. However, it should be clear that this will cause no loss of generality.

can be seen by setting $y := A[x']$ in the definition, for which there exists a valid decoding $z \in \mathbb{F}_2^n$. As $(A[x], y)$ are (e_0, e_1)-close, the definition implies that (x, z) must be (e'_0, e'_1)-close. Moreover, $(A[x'], y)$ are $(0, 0)$-close and thus, (e_0, e_1)-close, which implies that (z, x') must be (e'_1, e'_0)-close. Thus by the triangle inequality, (x, x') must be $(e'_0 + e'_1, e'_0 + e'_1)$-close.

Now, observe that for all i, (x_i, x_{i+1}) are $(e'_0+e'_1, e'_0+e'_1)$-far, and hence, their encodings must be (e_0, e_1)-far, by the fact we just mentioned. In particular this implies that $A[x_t]$ must have weight at least $t(e_0 + 1)$, which must be trivially upper bounded by m. Hence it follows that $(e_0+1)/(e'_0+e'_1+1) \leq m/d$. Similarly we can also show that $(e_1 + 1)/(e'_0 + e'_1 + 1) \leq m/d$.

C.2 Proof of Lemma 3

Let $x \in \mathbb{F}_2^n$ be chosen uniformly at random among vectors of weight d. Randomly flip $e'_1 + 1$ of the bits on the support of x to 0, and denote the resulting vector by x'. Using the partial ordering $\prec$ in the proof of the last lemma, it is obvious that $x' \prec x$, and hence, $A[x'] \prec A[x]$. Let b denote any disjunction of a number of coordinates in x and b' the same disjunction in x'. We must have

$$\Pr[b' = 0 | b = 1] \leq \frac{e'_1 + 1}{d},$$

as for b to be 1 at least one of the variables on the support of x must be present in the disjunction and one particular such variable must necessarily be flipped to bring the value of b' down to zero. Using this, the expected Hamming distance between $A[x]$ and $A[x']$ can be bounded as follows:

$$\mathbb{E}[\mathsf{dist}(A[x], A[x'])] = \sum_{i \in [m]} \mathbb{1}(A[x]_i = 1 \wedge A[x']_i = 0) \leq \frac{e'_1 + 1}{d} \cdot m,$$

where the expectation is over the randomness of x and the bit flips. Fix a particular choice of x' that keeps the expectation at most $(e'_1 + 1)m/d$. Now the randomness is over the possibilities of x, that is, flipping up to e'_1+1 zero coordinates of x' randomly. Denote by $\mathcal{X}$ the set of possibilities of x for which $A[x]$ and $A[x']$ are $\frac{(e'_1+1)m}{\epsilon d}$-close, and by $\mathcal{S}$ the set of all vectors that are monotonically larger than x' and are $(e'_1 + 1)$-close to it. Obviously, $\mathcal{X} \subseteq \mathcal{S}$, and, by Markov's inequality, we know that $|\mathcal{X}| \geq (1 - \epsilon)|\mathcal{S}|$.

Let z be any valid decoding of $A[x']$, Thus, (x', z) must be (e'_0, e'_1)-close. Now assume that $e_1 \geq \frac{(e'_1+1)m}{\epsilon d}$ and consider any $x \in \mathcal{X}$. Hence, $(A[x], A[x'])$ are (e_0, e_1)-close and (x, z) must be (e'_0, e'_1)-close by Definition 1. Regard x, x', z as the characteristic vectors of sets $X, X', Z \subseteq [n]$, respectively, where $X' \subseteq X$. We know that $|X \setminus Z| \leq e'_1$ and $|X \setminus X'| = e'_1 + 1$. Therefore,

$$|(X \setminus X') \cap Z| = |X \setminus X'| - |X \setminus Z| + |X' \setminus Z| > 0, \tag{1}$$

and z must take at least one nonzero coordinate from $\mathsf{supp}(x) \setminus \mathsf{supp}(x')$.

Now we construct an $(e_1'+1)$-hypergraph H as follows: The vertex set is $[n] \setminus \mathsf{supp}(x')$, and for every $x \in \mathcal{X}$, we put a hyperedge containing $\mathsf{supp}(x) \setminus \mathsf{supp}(x')$. The density of this hypergraph is at least $1-\epsilon$, by the fact that $|\mathcal{X}| \geq (1-\epsilon)\mathcal{S}$. Now Lemma 18 implies that H has a matching of size at least

$$t := \frac{(1-\epsilon)(n-d+1)}{(e_1'+1)^2}.$$

As by (1), $\mathsf{supp}(z)$ must contain at least one element from the vertices in each hyperedge of this matching, we conclude that $|\mathsf{supp}(z) \setminus \mathsf{supp}(x')| \geq t$, and that $e_0' \geq t$.

C.3 Proof of Lemma 4

For integers $a > b > 0$, we use the notation $V(a,b)$ for the volume of a Hamming ball of radius b in $\mathbb{F}_2^a$. It is given by

$$V(a,b) = \sum_{i=0}^{b} \binom{a}{i} \leq 2^{ah(b/a)},$$

where $h(\cdot)$ is the binary entropy function, and thus

$$\log V(a,b) \leq b \log \frac{a}{b} + (a-b) \log \frac{a}{a-b} = \Theta(b \log(a/b)).$$

Also, denote by $V'(a,b,e_0,e_1)$ the number of vectors in $\mathbb{F}_2^a$ that are (e_0,e_1)-close to a fixed b-sparse vector. Obviously, $V'(a,b,e_0,e_1) \leq V(b,e_0)V(a-b,e_1)$. Now consider any (wlog, deterministic) reconstruction algorithm D and let X denote the set of all vectors in $\mathbb{F}_2^n$ that it returns for some noiseless encoding; that is,

$$X := \{x \in \mathbb{F}_2^n \mid \exists y \in \mathcal{B}, x = D(A[y])\},$$

where $\mathcal{B}$ is the set of d-sparse vectors in $\mathbb{F}_2^n$. Notice that all vectors in X must be $(d+e_0')$-sparse, as they have to be close to the corresponding "correct" decoding. For each vector $x \in X$ and $y \in \mathcal{B}$, we say that x is *matching* to y if (y,x) are (e_0',e_1')-close. A vector $x \in X$ can be matching to at most $v := V'(n, d+e_0', e_0', e_1')$ vectors in $\mathcal{B}$, and we upper bound $\log v$ as follows:

$$\log v \leq \log V(n-d-e_0', e_1') + \log V(d+e_0', e_0') = O(e_1' \log((n-d-e_0')/e_1')) + d + e_0',$$

where the term inside $O(\cdot)$ is interpreted as zero when $e_1' = 0$. Moreover, every $y \in \mathcal{B}$ must have at least one matching vector in X, namely, $D(A[y])$. This means that $|X| \geq |\mathcal{B}|/v$, and that

$$\log |X| \geq \log |\mathcal{B}| - \log v \geq d \log(n/d) - d - e_0' - O(e_1' \log((n-d-e_0')/e_1')).$$

Finally, we observe that the number of measurements has to be at least $\log |X|$ to enable D to output all the vectors in X.

C.4 Proof of Theorem 8

Index the coordinates of S by the elements of $\mathbb{F}_2^t$ and denote the ith coordinate by S_i. Let Y be any random variable with min-entropy at least $t+k'$ distributed on $\mathbb{F}_2^{t+k'}$. Define an information theoretic test $T\colon \mathbb{F}_2^\ell \times \mathbb{F}_2^t \to \mathbb{F}_2$ as follows: $T(x,i) = 1$ if and only if $x \in S_i$. Observe that $\Pr[T(Y)=1] \leq \mathsf{wgt}(S)2^{-(t+k')} = \rho(S)2^{\ell-k'}$, and that for every vector $w \in (\mathbb{F}_2^\ell)^{2^t}$, $\Pr_{i\sim\mathcal{U}_t}[T(w_i,i)=1] = \mathsf{Agr}(w,S)$. Now let the random variable $X = (X_1,\ldots,X_{2^t})$ be uniformly distributed on the codewords in $\mathsf{LIST}_{\mathcal{C}}(S, \rho(S)2^{\ell-k'}+\epsilon)$ and $Z \sim \mathcal{U}_t$. Thus, from Definition 5 we know that $\Pr_{X,Z}[T(X_Z,Z)=1] > \rho(S)2^{\ell-k'}+\epsilon$. As the choice of Y was arbitrary, this implies that T is able to distinguish between the distribution of (Z,X) and any distribution on $\mathbb{F}_2^{t+\ell}$ with min-entropy at least $t+k'$, with bias greater than ϵ, which by the definition of condensers implies that the min-entropy of X must be less than k, or $|\mathsf{LIST}_{\mathcal{C}}(S, \rho(S)2^{\ell-k'}+\epsilon)| < 2^k$.

C.5 Proof of Theorem 9

Denote by $\mathcal{M}$ the adjacency matrix of the codeword graph of $\mathcal{C}$ and by M the number of its rows. It immediately follows from the construction that $M = TL$. Moreover, notice that the Hamming weight of each column of $\mathcal{M}$ is exactly T. Let $x \in \mathbb{F}_2^N$ and denote by $y \in \mathbb{F}_2^M$ its encoding, i.e., $y := \mathcal{M}[x]$, and by $\hat{y} \in \mathbb{F}_2^M$ a *received word*, or a *noisy* version of y. The encoding of x can be schematically viewed as follows: The coefficients of x are assigned to the left vertices of the codeword graph and the encoded bit on each right vertex is the bitwise OR of the values of its neighbors. The coordinates of x can be seen in one-to-one correspondence with the codewords of $\mathcal{C}$. Let $X \subseteq \mathcal{C}$ be the set of codewords corresponding to the support of x. The coordinates of the noisy encoding $\hat{y}$ are indexed by the elements of $[T] \times [L]$ and thus, $\hat{y}$ naturally defines a mixture $S = (S_1,\ldots,S_T)$ over $[L]^T$, where S_i contains j iff $\hat{y}$ at position (i,j) is 1. Observe that $\rho(S)$ is the relative Hamming weight (denoted below by $\delta(\cdot)$) of $\hat{y}$; thus,

$$\rho(S) = \delta(\hat{y}) \leq \delta(y) + p \leq \frac{D}{L} + p = \gamma + p,$$

where the last inequality comes from the fact that the relative weight of each column of $\mathcal{M}$ is exactly $1/L$ and that x is D-sparse. Furthermore, from the assumption we know that the number of false negatives in the measurement is at most $\nu TL/D = \nu T/\gamma$. Therefore, any codeword in X must have agreement at least $1-\nu/\gamma$ with S. This is because S is indeed constructed from a mixture of the elements in X, modulo false positives (that do not decrease the agreement) and at most $\nu T/\gamma$ false negatives each of which can reduce the agreement by at most $1/T$.

Accordingly, we consider a decoder which simply outputs a binary vector $\hat{x}$ supported on the coordinates corresponding to those codewords of $\mathcal{C}$ that have agreement larger than $1-\nu/\gamma$ with S. Clearly, the running time of the decoder is linear in the size of the measurement matrix. By the discussion above, $\hat{x}$

must include the support of x. Moreover, Theorem 8 applies for our choice of parameters, implying that the Hamming weight of $\hat{x}$ must be less than K.

C.6 Proof of Corollary 10

For simplicity we assume that $N = 2^n$ and $D = 2^d$ for positive integers n and d. However, it should be clear that this restriction will cause no loss of generality and can be eliminated with a slight change in the constants behind the asymptotic notations.

We instantiate the parameters of Theorem 9 using an optimal strong extractor. If $\nu = 0$, we choose γ, ϵ small constants such that $\gamma + \epsilon < 1 - p$. Otherwise, we choose $\gamma := \sqrt[3]{\nu}$, which makes $\nu/\gamma = \sqrt[3]{\nu^2}$, and $\epsilon < 1 - p - \sqrt[3]{\nu} - \sqrt[3]{\nu^2}$. (One can easily see that the right hand side of the latter inequality is positive for $\nu < \nu_0$). Hence, the condition $p + \nu/\gamma < 1 - \epsilon - \gamma$ required by Theorem 9 is satisfied. Let $r = 2\log(1/\epsilon) + O(1) = O(1)$ be the entropy loss of the extractor for error ϵ, and set up the extractor for min-entropy $k = \log D + \log(1/\gamma) + r$, which means that $K = 2^k = O(D)$ and $L = 2^\ell = D/\gamma = O(D)$. Now we can apply Theorem 9 and conclude that the measurement matrix is $(pM, (\nu/D)M, O(D), 0)$-correcting. The seed length required by Ext is $t \leq \log n + 2\log(1/\epsilon) + O(1)$, which gives $T = 2^t = O(\log N)$. Therefore, the number of measurements will be $M = TL = O(D \log N)$.

C.7 Proof of Corollary 11

We will use the notation of Theorem 9 and apply it using an optimal strong lossless condenser. Set up the condenser with error $\epsilon := \frac{1}{2}\delta/(1+\delta)$ and min-entropy k such that $K = 2^k = D/(1-2\epsilon)$. As the error is a constant, the overhead and hence L/K will also be a constant. The seed length is $t = \log(n/\epsilon) + O(1)$, which makes $T = O(\log N)$. As $L = O(D)$, the number of measurements will be $M = TL = O(D \log N)$, as desired. Moreover, note that our choice of K will imply that $K - D = \delta D$. Thus we only need to choose p and ν appropriately to satisfy the condition $(p + \gamma)L/K + \nu/\gamma < 1 - \epsilon$, where $\gamma = D/L = K/(L(1+\delta))$ is a constant, as required by the lemma. Substituting for γ, we will get the condition $pL/K + \nu L/(K(1+\delta)) < \delta/(1+\delta)$, which can be satisfied by choosing p and ν to be appropriate positive constants.

D A Combinatorial Lemma

For a positive integer $c > 1$, define a c-hypergraph as a tuple (V, E), where V is the set of vertices and E is the set of hyperedges and every $e \in E$ is a subset of V of size c. The degree of a vertex $v \in V$, denoted by $\mathsf{deg}(v)$, is the size of the set $\{e \in E \colon v \in E\}$. Note that $|E| \leq \binom{|V|}{c}$ and $\mathsf{deg}(v) \leq \binom{|V|}{c-1}$. The *density* of the hypergraph is given by $|E|/\binom{|V|}{c}$. A *vertex cover* on the hypergraph is a subset of vertices that contains at least one vertex from every hyperedge. A *matching* is a set of pairwise disjoint hyperedges. It is well known that any dense hypergraph must have a large matching. Below we reconstruct a proof of this claim.

Proposition 17. *Let H be a c-hypergraph such that every vertex cover of H has size at least k. Then H has a matching of size at least k/c.*

Proof. Let M be a maximal matching of H, i.e., a matching that cannot be extended by adding further hyperedges. Let C be the set of all vertices that participate in hyperedges of M. Then C has to be a vertex cover, as otherwise one could add an uncovered hyperedge to M and violate maximality of M. Hence, $c|M| = |C| \geq k$, and the claim follows. □

Lemma 18. *Let $H = (V, E)$ be a c-hypergraph with density at least $\epsilon > 0$. Then H has a matching of size at least $\frac{\epsilon}{c^2}(|V| - c + 1)$.*

Proof. For every subset $S \subseteq V$ of size c, denote by $\mathbb{1}(S)$ the indicator value of S being in E. Let C be any vertex cover of H. Denote by $\mathcal{S}$ the set of all subsets of V of size c. Then we have

$$\epsilon\binom{|V|}{c} \leq \sum_{S \in \mathcal{S}} \mathbb{1}(S) \leq \sum_{v \in C} \mathsf{deg}(v) \leq |C|\binom{|V|}{c-1}.$$

Hence, $|C| \geq \epsilon(n - c + 1)/c$, and the claim follows using Proposition 17. □

m	e_0	e_1	e_0'	Det/ Rnd	Rec. Time
$O(d \log n)$	αm	$\Omega(m/d)$	$O(d)$	Rnd	$O(mn)$
$O(d \log n)$	$\Omega(m)$	$\Omega(m/d)$	δd	Rnd	$O(mn)$
$O(d^{1+o(1)} \log n)$	αm	$\Omega(m/d)$	$O(d)$	Det	$O(mn)$
$d \cdot \mathsf{quasipoly}(\log n)$	$\Omega(m)$	$\Omega(m/d)$	δd	Det	$O(mn)$
$d \cdot \mathsf{quasipoly}(\log n)$	αm	$\Omega(m/d)$	$O(d)$	Det	$\mathsf{poly}(m)$
$\mathsf{poly}(d)\mathsf{poly}(\log n)$	$\mathsf{poly}(d)\mathsf{poly}(\log n)$	$\Omega(e_0/d)$	δd	Det	$\mathsf{poly}(m)$

Table 1. A summary of constructions in this paper. The parameters $\alpha \in [0, 1)$ and $\delta \in (0, 1]$ are arbitrary constants, m is the number of measurements, e_0 (resp., e_1) the number of tolerable false positives (resp., negatives) in the measurements, and e_0' is the number of false positives in the reconstruction. The fifth column shows whether the construction is deterministic (Det) or randomized (Rnd), and the last column shows the running time of the reconstruction algorithm.